JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 37, 203-218 (2021)
DOI: 10.6688/JISE.202101_37(1).0013

Acoustic Anomaly Detection Using Multilayer Neural Networks and Semantic Pointers

CHE-JUI CHANG AND SHYH-KANG JENG
Department of Electronic Engineering
National Taiwan University
Taipei, 106 Taiwan
E-mail: {r05942055; skjeng}@ntu.edu.tw

In this paper we provide an auditory processing system with higher biological plausibility than previous studies to solve the problem of acoustic anomaly detection in household environment. First, in the proposed system the log filter bank is adopted for extracting audio features, simulating the function of the peripheral auditory system (outer, middle and inner ears to auditory nerves). Next, we use multi-layer neural networks to imitate auditory cortex in human brains, in order to extract abstract semantic contents. Then, a semantic pointer architecture model is used to imitate prefrontal cortex, basal ganglia, and thalamus, in which the anomaly is detected using symbol-like rules. Compared with other anomaly detection methods with different biological plausibility in performance, our proposed method gets the best result on the testing set, with 0.956 AUC. Meanwhile, it takes less computational time to detect the anomaly. Hence, it is suitable for detecting acoustic anomalies in real-world cases.

Keywords: acoustic anomaly detection, biological plausibility, household robot, humanoid auditory processing, neural networks, semantic pointers

1. INTRODUCTION

Detection of acoustic anomaly, such as scream and alarm in daily life, is an important area and has many practical applications as addressed in [1]. Surveys of anomaly detection, or more generally, novelty detection, can be found in some literatures [2, 3]. Typically, most anomaly detection problems have large amount of normal data, but little or rare abnormal data, which makes anomaly detection different from two-class or multi-class classification. Based on the surveys [2, 3], a general method to deal with this problem is one-class training, which uses a classifier trained only on normal data, and outputs an anomaly score for each input data. The score can be obtained from a distance, similarity, probability, density, or reconstruction error. Many one-class approaches to acoustic anomaly detection have been proposed and reviewed in [1].

However, most of the studies are based on purer mathematical manipulation and are less related to a high-performance machine in the nature, the auditory system of us human. In fact, million years of evolution has provided us an adaptive and robust auditory system for anomaly detection, protecting us from being attacked. It is safe to say that human auditory system is a perfect subject for researchers to investigate, or even imitate. Inspired by this idea, in this study we aim to build a system with higher biological plausibility (BP) for anomaly detection. Compared with previous works in [2, 3], our aiming method is expected to be much more akin to the human auditory system. To quantify the differences,

Received November 7, 2018; revised January 2 & February 10, 2019; accepted March 11, 2019.
Communicated by Chung-Hsien Wu.

we use a qualitative measure, BP value, to roughly estimate a method's proximity to the corresponding human system. Those methods using pure statistical calculations without taking into considerations the relation to biological perception system will be assigned a very low, or zero, BP value. On the contrary, systems with every aspect corresponding to the process of a biological one can be assigned a very high BP value (with the top BP value: +++++, where the more plus signs means the higher BP value). In this paper, encouraged by previous work [4], we propose a new system imitating the human auditory system for detecting abnormal sound. The simplified auditory system uses a neural network and a semantic pointer architecture [5] to imitate the human auditory cortex and the prefrontal cortex. Though it may not be close to the nature neural system in every aspect, its biological plausibility is still much higher than other methods in literature, and believed to have a BP value of +++.

The term "semantic pointer (SP)" was coined by Chris Eliasmith in University of Waterloo [5], and is considered high-level cognitive neural representations. Semantic pointers, also called semantic vectors in this paper, carry partial semantic content in cognitive functions and their corresponding vector representations can be transmitted through neural sub-systems. Mathematically, a set of semantic pointers is actually a set of high-dimensional vectors approximately orthogonal to each other. These high-level vectors can be generated or extracted from neural networks. A common and easy method to do this is to use an autoencoder: the encoding part generates semantic pointers from low-level features and the decoding part reconstructs the low-level vectors from semantic pointers. These semantic pointers, on the other hand, can be manipulated in a semantic pointer architecture (SPA), which is a large-scale cognitive neural model consisting of populations (or ensembles) of spiking neurons. The neuron ensemble, which acts as a working unit in SPA, is used to represent semantic vector states by encoding them into spiking neural signals. According to the principles of Neural Engineering Framework proposed by Chris Eliasmith in [5], neuron ensembles can also connect to each other, allowing signals to transmit and processed.

One important advantage of SPA is that its idea is in fact a kind of vector symbolic architecture [6], if not considering its underneath neural models. The orthogonal semantic pointers, or vectors, can be bound by tags (also semantic pointers) or added together to create another new semantic pointer. For example, in the notation system adopted in [5], the sentence, "a dog chases after a boy," can be simply represented as a semantic pointer P by the composition of semantic pointers in the vector form: "P = SUBJECT⊛DOG + VERB⊛CHASE + OBJECT⊛BOY." In this example, the sentence has a subject, DOG; a verb, CHASE; and an object, BOY, so the semantic pointer P is composed of three binding pairs of tag SP and content SP. The operator "⊛" is related to taking the circular convolution over its left and right operands. Besides, we may ask questions about the tags, like "What is the verb?" and get the answer by unbinding the P vector with the inverse of the "⊛" operation corresponding to the query tag, such as "P⊛~VERB", and the result will be a SP with a vector form close to that of CHASE. A similar idea related to the vector symbolic architecture and semantic pointers is the hyperdimensional computing [7], which uses high-dimensional binary vectors to store and represent semantic information. In addition, hyperdimensional computing has similar vector computation rules to compose or decompose vectors. Recently, it has been applied for visual question answering [8].

To use spiking neurons for cognitive process of SPA models, the open-source neural

simulator Nengo[1], which is also developed by the Eliasmith team, is usually used. Here Nengo represents "Neural ENGineering Object". It provides lucid API documentations and practical examples for users. In Nengo, all neurons, which process signals and semantic pointers, are with simplified models simulating production of neural spikes in biological neural cells.

In this paper, we will describe our simplified model for the human auditory system in Section 2. The method to use this simplified model for acoustic anomaly detection is proposed in Section 3. The data preparation and simulation setup are described in Section 4. Results and discussions are presented in Section 5. Finally, Section 6 concludes this paper.

2. SIMPLIFIED AUDITORY SYSTEM

In human auditory system [9-11], audio signal processing starts from the peripheral auditory system (PAS), passing signals through outer, middle and inner ears to auditory nerves. Neural signals from auditory nerves then pass through ascending auditory pathways and reach the auditory cortex. Next, signals are processed in the auditory cortex and transmitted to the prefrontal cortex for anomaly detection.

A comparison of the signal flow in the human auditory system and in our simplified system is shown in Fig. 1. More descriptions for each module are given in the following subsections.

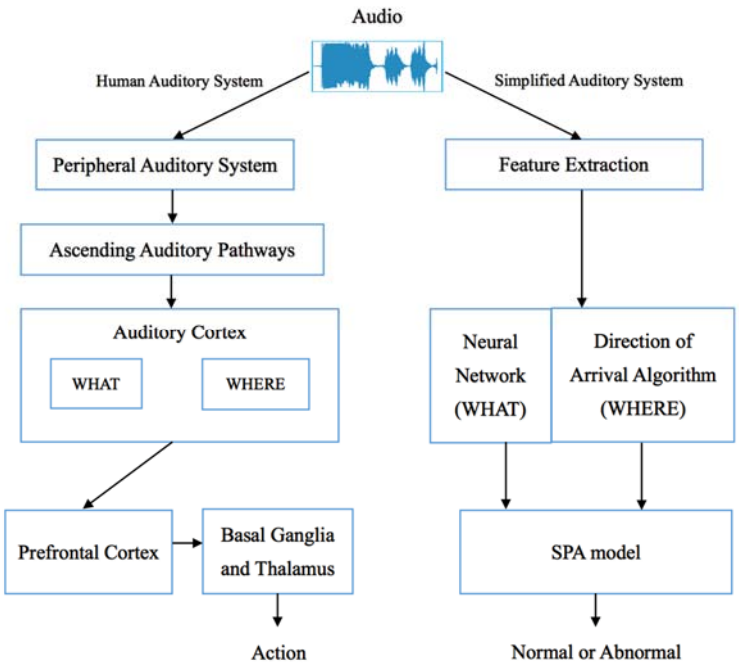

Fig. 1. Human auditory system and our simplified artificial auditory system.

[1] https://www.nengo.ai/

2.1 Peripheral Auditory System

The peripheral auditory system (PAS) [9] is composed of the outer, middle, and inner ear and transmits audio signal to auditory nerves. It can be imitated by some feature extraction methods. One is the IPEM method, which transforms input signals into features by using the auditory peripheral module of Toolbox[2] in MATLAB [12]. Here IPEM stands for the research group, Institute for Psychoacoustics and Electronic Music in Ghent University. The method can better simulate the PAS because of its 40-channel filter banks with linear filters and non-linear hair cell models. These band pass filter banks are imitation of the cochlea model, processing audio signals to IPEM features. Other popular feature extraction methods such as Mel-Frequency Cepstrum Coefficients (MFCC) [13] and log filter bank method [13] (MFCC without the final discrete cosine transform stage) can also be treated in some way as artificial processes imitating the PAS: the triangular-shaped overlapping filters in these two methods are loosely related to the cochlea in the human auditory system, and are thus with some biological plausibility value.

Although IPEM simulates the PAS better and has higher BP value, it is inferior to the other two methods in running time efficiency. Besides, we found that the MFCC method works not well in our system, either (see Section 3.1). Thus, in our proposed system, we adopt the log filter bank method to extract features from audio signals.

2.2 Ascending Auditory Pathways

In a human auditory system, signals processed by peripheral auditory system are then transmitted through the ascending auditory pathways [10, 11] to auditory cortex. The binaural pathways begin from auditory nerves, passing upwards through some nerve nucleus, like Ventral Cochlea Nucleus and Superior Olivary Complex, and at last arriving at the auditory cortex. The ascending auditory pathways serve as a major function for the direction of arrival estimation for human, through which the interaural time difference and interaural level difference are calculated and integrated for the estimation.

Though previous studies [12] have provided ample resolutions and implementations to the ascending auditory pathways, the major function of the pathways is not necessary for the processing of anomaly detection. Thus, we ignore the ascending auditory pathways in our proposed system: features extracted from the artificial peripheral auditory system are directly sent to our artificial auditory cortex, without any process corresponding to the ascending auditory pathways.

2.3 Auditory Cortex

Auditory cortex [11] is composed of some sub-areas like the primary auditory cortex (A1), belt, and parabelt. When the neural signal arrives at these sub-areas, it is processed and split into two processing channels. One is for the auditory WHAT process and the other is for the auditory WHERE process. The WHAT channel is in charge of recognizing semantic contents of the incoming audio signal, while the WHERE channel deals with sound localization and determines the direction of the sound source. Signals passing

[2] https://www.ugent.be/lw/kunstwetenschappen/en/research-groups/musicology/ipem/finishedprojects/ipem-toolbox.htm

through these two channels will reach the prefrontal cortex for even higher and abstractive processes.

In this study, the process of the WHAT channel is imitated by a multilayer artificial neural network, as the encoding part of an autoencoder, which can extract semantic information from the input feature vectors and build a semantic pointer from the output layer (see Section 3.2). The semantic pointer (SP) uses a symbol (such as DOG) and a numeric high-dimensional vector representing the firing states of all neurons in a neural population (ensemble). The hidden layers in the neural network loosely imitate the sub-areas in the auditory cortex. Moreover, since the auditory cortex in human brain can adapt to the surroundings through learning, the weights in the neural network model can also be trained and tuned to simulate the neural plasticity. On the other hand, we adopt methods in [14] for the process of the WHERE channel. The output vectors from both channels will be fed into our SPA model.

2.4 Prefrontal Cortex

The signals outputted from the two channels in auditory cortex are sent to the prefrontal cortex (PFC) [11] for higher cognitive level semantic processing. In this study, we use artificial spiking neural models to build our semantic pointer architecture (SPA) model for the PFC. The SPA are system schemes dealing with the semantic pointers using circular convolution (binding) and addition. In the PFC model, we bind vectors, from the WHAT channel and the WHERE channel, to a new semantic pointer. Based on the single semantic pointer, our artificial PFC is able to continuously answer if the input sound is anomaly and where the sound comes from. Since the semantic pointer is related to firing states of neurons, and the operations of binding and addition can be implemented based on the transformation of neural signals [5], our PFC model is also in some sense biologically plausible.

2.5 Basal Ganglia and Thalamus

Although the basal ganglia and the thalamus are not parts of the auditory system, they play important roles for motion control, action selection, and transmission control [5]. In our system, we use artificial basal ganglia and thalamus, both implemented in SPA model, to imitate the function of action selection in the basal ganglia and transmission control in the thalamus. Since the basal ganglia will select the action with higher utility, we set the utility function as the anomaly score so that the basal ganglia model can determine whether the input sound is anomalous.

3. PROPOSED METHOD

Our proposed method and architecture are shown in Fig. 2. For the auditory WHAT process, first, we extract feature vectors from .wav audio files. Then, a well-trained 4-layer neural network model imitating major sub-areas of the auditory cortex is applied to construct high-level semantic pointers. The semantic pointers, containing the compressed semantic information of sound content, are sent to the SPA model. The artificial basal ganglia receive the semantic pointers and then detects the anomaly by comparing the corresponding utility values.

In addition to the basal ganglia for anomaly detection, we build a structure for question answering in the SPA model. The outcome semantic pointer, which is NORMAL or ABNORMAL, from the basal ganglia process and the angle semantic pointer from the auditory WHERE process are bound together to create a new semantic pointer. The new created vector, containing normality and angle information, is then used to answer both of the questions, Q1 and Q2, shown in Fig. 2.

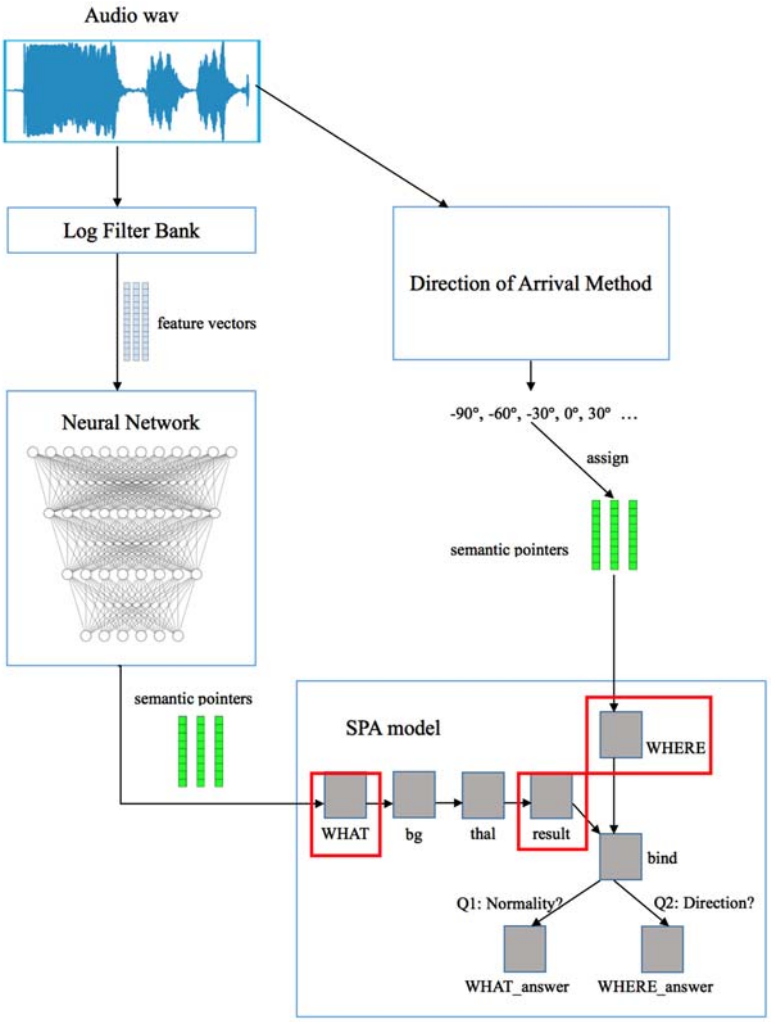

Fig. 2. Proposed system structure for acoustic anomaly detection.

Fig. 2 proposed a system structure for acoustic anomaly detection. Note that there are two semantic pointer inputs, from WHAT and WHERE channel, for SPA model. In SPA model, each gray block is a neuron ensemble representing a semantic pointer state. The description of labels of the ensembles are as follows: WHAT: an ensemble representing the input SP state from neural network model; WHERE: an ensemble representing the

input SP state from WHERE channel; bg: basal ganglia area which computes the similarity between the semantic pointer from WHAT ensemble and NORMAL semantic vector, and selects an action; thal: thalamus, for transmitting signals to output neural ensemble; result: an ensemble representing the SP state for result of anomaly detection; bind: an ensemble representing state of the new SP which is a binding of the SPs from two channels with tags; WHAT_answer: an ensemble representing SP state for the answer to the question of normality; WHERE_answer: an ensemble representing SP state for the answer to question of the direction of source.

The following sections provide more detailed descriptions and procedures for each module.

3.1 Feature Extraction

After several tests, we found that the log filter bank method [13] is more accurate for validation data, while the features extracted by IPEM method [12] perform a little better for testing data. Although previous work [14] has provided new algorithms to do filter bank calculation in every channel in parallel and improve the computational time for extracting IPEM features, using IPEM method, however, still takes much more time for computation than the log filter bank method. On the other hand, the MFCC method [13], using discrete cosine transform followed by log filter banks, has the poorest performance in our tests because it concentrates most energy on the first feature, which is believed to be more difficult for reconstruction by the autoencoder (see Section 3.2) and makes the loss value higher during training. With these concerns, in this paper, the log filter bank method is adopted for feature extraction.

3.2 Neural Network Model

The main purpose of our neural network model is to transform feature vectors into semantic vectors as shown in Fig. 2, so the model will take the extracted features as input and generate semantic pointers. To train the model, we fold the network to form an autoencoder and adopt one-class learning method to learn compressed representations of normal data. We also expect that the inference model, which is the encoder part of the autoencoder, can concentrate on normal vectors and separate abnormal vectors in the semantic space so that anomaly detection can be conducted by comparing similarities. To do this, first of all, the two symbolic labels, NORMAL and ABNORMAL, are generated as orthogonal and random vectors in the semantic space. Next, for one-class training, we use only NORMAL as label for normal data in the neural network. Then, in order to cluster normal data, one important training target is to transform the normal data into semantic vectors that are close to the NORMAL vector in semantic space. Hence, our training method, different from conventional methods that use only reconstruction loss to learn compressed representations, gives the autoencoder network an additional loss component, $loss_{sp}$, so that the compressed semantic vectors of normal data will become closer to the NORMAL vector in the semantic space. Thus, we define the loss function of our proposed neural network model as Eq. (1) and minimize the loss value to meet this additional training goal.

$$loss = loss_{rec} + \lambda \cdot loss_{sp} \tag{1}$$

Here $loss_{rec}$ stands for the reconstruction loss of the autoencoder, and $loss_{sp}$ represents the distance measure between two semantic pointers, namely, the generated semantic vector and the NORMAL vector. Also, λ is a constant in Eq. (1). If the λ value is close to 0, the model will take only reconstruction loss into account, but the generated vectors from the encoder may not be close to our expected NORMAL vector. On the contrary if λ is very large, $loss_{sp}$ will be dominant and the generated vectors will lose their diversity. In our experiment, we set λ to be 0.4 and the outcome of anomaly detection has an F1-score of 0.86, which is 1 percent better than a multilayer perceptron autoencoder [1] with only reconstruction loss.

After some tests, a total of four hidden layers are used and cascaded in our neural network model, as shown in Fig. 2. We use ReLU as the activation function except for the encoder output layer and the decoder output layer. A hyperbolic tangent function is adopted as the activation function of encoder output layer to ensure the output to be between −1 and 1, and a linear function is used in the decoder output layer. For training of this neural network model, the two loss functions in Eq. (1) are defined as mean square errors, respectively, and the method of backpropagation [15] is applied to update the weights.

As the training finishes, we use only the encoder part of the autoencoder to receive input feature vectors and output the corresponding semantic pointers for auditory WHAT channel.

3.3 SPA Model

As the neural network is well-trained, our SPA model will calculate the cosine similarities between the generated semantic pointers (SPs) and NORMAL, and detect anomalies by comparing them with a given threshold. Semantic pointers obtained from our artificial auditory cortex model is fed to the "WHAT" ensemble. Then, the basal ganglia (ensemble "bg") receives the SP from "WHAT", calculates the similarity of input SP and NORMAL. It will also compare the similarity with a threshold. If the similarity is large enough, it will take actions through thalamus (ensemble "thal") and assign a NORMAL label to ensemble "result". On the other hand, if the similarity does not exceed the threshold value, the sound is far from normal, so the ABNORMAL vector will be assigned to ensemble "result". The basal ganglia action for our proposed system is defined through Eqs. (2) and (3), using the notation system in [5].

$$dot(WHAT, NORMAL) \rightarrow result = NORMAL, \tag{2}$$

$$threshold \rightarrow result = ABNORMAL. \tag{3}$$

In addition to anomaly detection in our SPA model, we take advantage of semantic pointer representations to answer questions. For our model, the outcome vector of anomaly detection, represented by ensemble "result", is bound with SP from the WHERE channel, represented by the ensemble "WHERE", to create a new SP, represented by "bind" ensemble, as shown in Eq. (4).

$$bind = result \circledast \text{Normality} + \text{WHERE} \circledast \text{Direction} \tag{4}$$

The bound SP contains information from both of WHAT and WHERE channels. For the question: "What is the label for normality," we parse the sentence and find the normality. Thus, we unbind the SP vector with Normality SP to get the answer SP, as Eq. (5). For the question: "Where does the sound source come from," we unbind the SP vector with Direction SP to get the answer SP, as Eq. (6).

$$WHAT\_answer = bind \circledast \sim \text{Normality}, \tag{5}$$

$$WHERE\_answer = bind \circledast \sim \text{Direction}. \tag{6}$$

4. DATA AND SIMULATION SETUP

4.1 Data and Preprocessing

We use the same dataset as that in a previous study [1], where most normal recordings come from household background sounds, extracted from PASCAL CHiME speech segmentation and recognition challenge [16]. The few anomaly recordings are selected and downloaded from an open-source website[3]. In the dataset of normal sounds, household background recordings consist of sounds like people talking, playing or watching television. On the other hand, the anomaly recordings consist of some types of abnormal sounds such as people falls, screams, alarms and fractures. There are around 100 and 70 minutes in our training and testing data sets, respectively. The sampling rate of all audio data is 16kHz.

For preprocessing, first, we divide audio clips into audio segments of 100 ms long and use log filter bank method with a frame size 25 ms and a shift 10 ms to extract features from every segment. Next, the extracted 9-frame feature vectors, with dimension 26, are concatenated to a vector with dimension 234. The vector will be standardized and then fed to the neural network as input. The reason why we use a 9-frame segment instead of one frame segment for anomaly detection is that the latter causes much more computation in the test phase as we don't need to detect the anomaly that frequently. Therefore, it is not an efficient method for real-time detection. Contrarily, the former method has a detection interval of 100 ms, which is more acceptable in our case.

4.2 Simulation Setups

After the neural network model is well trained, we use the Nengo package to simulate the spiking neural activities for the SPA model in Fig. 2. To do this, the input semantic pointers from two channels should be fed to the SPA model in a streaming mode. For the WHAT channel, we set the duration of each input semantic pointer to be 100 ms, which is exactly the time the original audio is segmented to predict anomaly. On the other hand, the duration is set to be 1 second for the WHERE channel. Moreover, for simplicity, we set the incoming angle directly in this experiment without processing the signal through the direction-of-arrival estimation system so that we can focus on the experiment on the SPA model, and examine the accuracy and exactitude of simulation results.

[3] https://freesound.org/

The direction setting in this work is mostly based on a previous study [14], which uses binaural microphone to classify the direction of arrival into 7 angles: 90°, 60°, 30°, 0°, 330°, 300°, and 270°. We use these 7 angles as all of our directions because people actually cannot distinguish very precisely the angle of arrival of the sound and an error of 15^{o} seems to be common. Thus, we set the difference between recognized angles to be 30^{o}. On the other hand, we think they are enough for our model to perform question answering tasks and verify its correctness. Moreover, if more angles are set in the experiment, it will be more difficult to generate orthogonal vectors in the semantic space and then assign to the angles. In this experiment, we select these 7 angles randomly as estimation results, and assign 7 orthogonal semantic pointers to these angles as input to our SPA model.

The simulation of the SPA model is in a streaming mode, so we can probe on any neuron ensemble and decode its output semantic pointer to observe the state change in its underlying high-dimensional vector during the simulation. For acoustic anomaly detection, we probe on the "result" ensemble, which represents the output state of the basal ganglia function. Cosine similarity between the semantic pointer from the "result" ensemble and the NORMAL vector is calculated as a score to measure the normality. To verify the result of basal ganglia function, we directly calculate cosine similarity between the vector obtained from the neural network model and the NORMAL vector before simulation. The pre-simulation and post-simulation results should be very close.

5. RESULTS

5.1 Post-Simulation: Anomaly Detection

We use basal ganglia to compare similarity and assign the detection outcome to "result" ensemble. The vector state of the probed neuron ensemble represents either NORMAL or ABNORMAL. Thus, for normal audio segments, the similarity between the outcome vector and the NORMAL vector should be larger or even close to the maximum value 1. We safely set a lower bound for the similarity value, 0.5, for normal class: data with its similarity score above the value is predicted as normal. Otherwise, it is abnormal.

In the test set, our proposed SPA model gets an F1-score of 0.86, which is almost the same as the pre-simulation result. This shows that a biologically plausible method can perform as well as the direct computation.

5.2 Post-Simulation: Question Answering

Once the basal ganglia detect the anomaly, we bind the semantic pointers from the two channels to create a new semantic pointer for question answering.

For the first question: "What is the label for normality?" we probe the "WHAT_answer" neuron ensemble to obtain its representing vector state and calculate the cosine similarities with the two label vectors, NORMAL and ABNORMAL. The anomaly score is defined as the difference of the similarities, and the result is shown in Fig. 3. We extract the first 20 seconds in the test set. Obviously, as the label becomes ABNORMAL, the corresponding anomaly score increases significantly, indicating that the sound is more likely to be an anomaly. The result also shows that our SPA model is able to store the

binding semantic pointer with enough content information and use the vector to answer the first question.

For the second question: "Where (Which direction angle) does the sound source come from?" the answer is represented by the "WHERE_answer" neuron ensemble. We calculate the similarities with all 7 angle semantic pointers and select the angle with the largest similarity. The result is shown in Fig. 4, where the estimated angles perfectly match the true angles in the first 20 seconds. This means that the SP vector contains not only the content information but also enough direction information so that our model is able to answer the second question correctly.

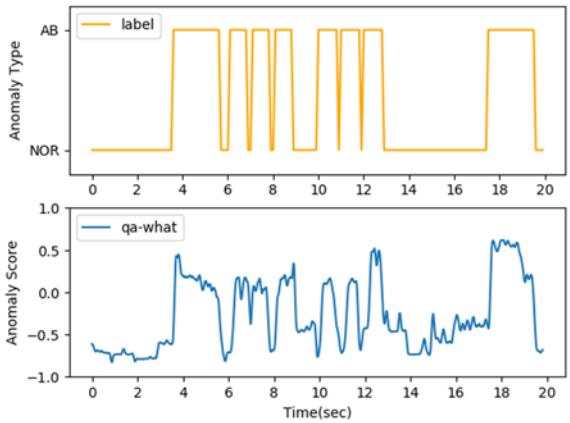

Fig. 3. Answering the first question (WHAT) of the first 20 seconds in test set with our SPA model.

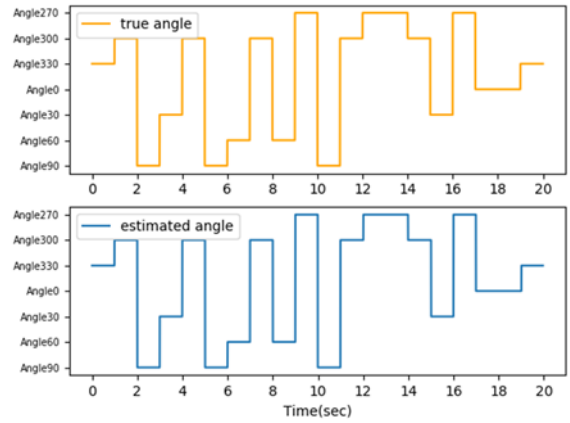

Fig. 4. Answering the second question (WHERE) of the first 20 seconds in test set with our SPA model.

5.3 Pre-Simulation: Effects of Different Sound Levels and SP Dimensions

Some parameters we are interested in in this model are the sound level and the dimension of semantic pointers. We would like to know how robust this model is to different

sound levels and how different SP dimensions affect the results. We increase and reduce the sound level, or volume, which results in sound levels from −6 dB to 12 dB with a 3dB increment. Note that the level 0 dB is corresponding to that of the recorded sound. We also use the models with different SP dimensions (32, 64, 128) to predict their capability of detecting anomalies. Assume that the anomaly score is one minus the similarity between the query SP and the NORMAL SP, we plot ROC curves for each case, and calculate their Area Under the Curve (AUC). The calculated AUC versus the sound level is depicted as Fig. 5.

We can see that the highest AUC (around 0.969) occurs at 3dB sound level and dimension of 32. Meanwhile, different SP dimensions seem to have little effect on the detection AUC when sound level is 0dB, 3dB and 6dB. As the sound level increases to 9dB or higher, the dimension of 32 still performs the best, 64 next, and 128 the last. However, as sound level decreases to −3dB or lower, higher SP dimension gets better accuracy, and the AUC with dimension of 32 at −6dB sound level drops dramatically to 0.70. These results show that our model are not robust enough to detect the acoustic anomaly at all sound levels, especially when the sound level is lower. Compared with the effect of different dimensions, different sound levels seem to have larger influence on the detection result. We also found that if the threshold is fixed, the model tends to predict more abnormal sound as the sound level increases, which makes both true positive rate and false positive rate higher.

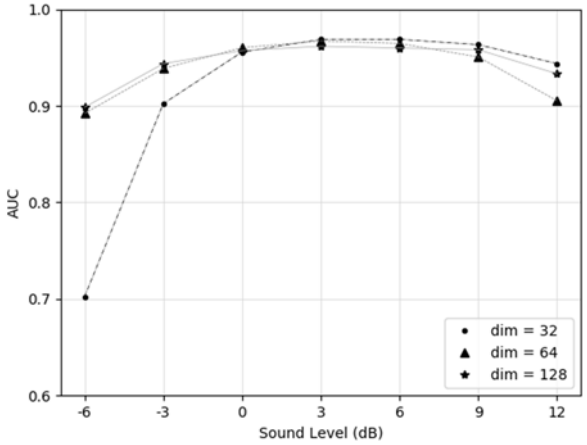

Fig. 5. AUC of different sound levels and different SP dimensions.

5.4 Comparison with Other Methods

In this section, we compare performances of different one-class anomaly detection models. **Rec-based** [1] stands for reconstruction-based methods. We trained a fully-connected autoencoder model by normal data, and then passed data features through the model to reconstruct the input features. The reconstruction error by comparing the original feature and the constructed vector is the anomaly score. Since neural networks were used for autoencoders, its biological plausibility is not bad. **LOF** [17] stands for local outlier factor, which is a distance-based anomaly detection method. Its algorithm first searches over k

nearest neighbors for the query data, and then the outlier factor is determined based on the *k* neighbors' density. The outlier factor is also regarded as an anomaly score. This is mostly a statistical approach, and the biological plausibility is small. **OCSVM** [18] stands for the one-class svm, which is a domain-based method, also with small biological plausibility. It uses kernel trick to project features to a high-dimensional space and separate the normal and abnormal data with a hyperplane in the high-dimensional space. The signed distance, positive for an inlier and negative for an outlier, to the separating hyperplane is calculated and the anomaly score of every data is defined as the negative signed distance.

The ROC results of all tested methods are shown in Fig. 6. Our proposed method (**NN-SP** in Fig. 6) with a dimension of 32 gets the best AUC result, 0.956. **Rec-based** has good results too, with AUC larger than 0.9. The **LOF** method gets the lowest AUC. What's worse is that it takes much longer time to find the neighboring points and predict the anomaly. Finally, **OCSVM** gets the secondly highest AUC in our experiments, with its AUC a little lower than that of our method. This can be interpreted as that both extremal cases of the biological plausibility, one is much higher and the other much lower, are possible to build accurate anomaly detection algorithms, although they are implemented by very different, but powerful, architectures. Nonetheless, algorithms with a higher biological plausibility are easier to integrate with systems in similar biologically plausible platforms for other tasks, such as the software robot, Spaun, described in [5]. This also indicates that the qualitative biological plausibility is better expressed for comparison among studies on cognitive systems.

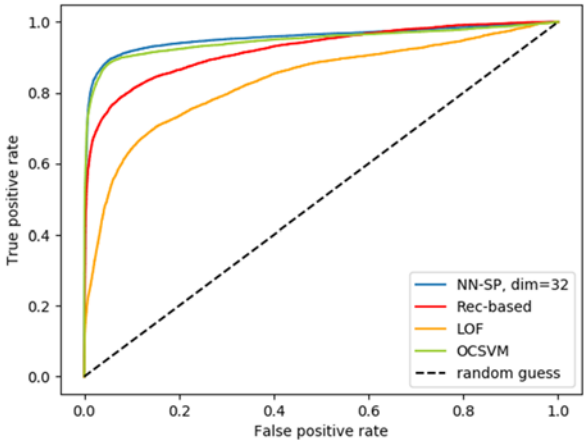

Fig. 6. ROC curves of different anomaly detection methods. The AUC's are 0.956, 0.921, 0.833, and 0.947 for this model (SP dimension = 32), **Rec-based**, **LOF**, and **OCSVM**, respectively.

6. CONCLUSIONS

We have proposed a simplified artificial auditory model on acoustic anomaly detection. Compared with previous works, we adopt an approach with higher biological plausibility (rough BP value: +++). Our proposed model includes imitations of Peripheral Auditory System (PAS), Auditory Cortex (AC), Prefrontal Cortex (PFC), and Basal Ganglia as

well as Thalamus, applying the log filter bank for the PAS, a 4-layer neural network for AC, and the Semantic Pointer Architecture (SPA) for PFC and Basal Ganglia. Besides, for the AC model, we proposed a new method to train the network using a two-loss autoencoder so that the generated semantic pointers will get close to the NORMAL vector and still maintain their diversity. The way we train the neural network enables us to detect the anomaly by comparing similarities of semantic pointers at the basal ganglia in the SPA model. Moreover, we take advantage of vector composition and decomposition to bind semantic pointers from two channels, fuse the information of normality and direction of arrival into the new semantic pointer, and use the stored semantic pointer for simple question-answering tasks.

We also compared the performance of our models with different hyper parameters and other methods with different biological plausibility. Among the often-applied anomaly detection methods, ours has the best result (0.956 AUC). Meanwhile, it takes less computational time to detect the anomaly. The performance of **OCSVM** (biological plausibility: nearly 0) is just a little inferior to that of our model. Hence, we conclude that cases with different biological plausibility are possible to result in different cognitive systems with comparable performances. For ease of comparative studies, a qualitative biological plausibility is better taken into considerations in developing, and discussed while presenting a cognition system.

In the future, we wish to integrate the model with other SPA modules for different tasks. A navigation module and a vision module are certainly most welcome so that the robot may move to the place where anomaly events happen. In addition, we look forward to further improving the performance, while not degrading the biological plausibility much.

REFERENCES

1. E. Marchi, F. Vesperini, F. Eyben, and S. Squartini, "A novel approach for automatic acoustic novelty detection using a dnoising autoencoder with bidirectional LSTM neural networks," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 1996-2000.
2. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection," *ACM Computing Surveys*, Vol. 41, 2009, pp. 1-6.
3. M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, Vol. 99, 2014, pp. 215-249.
4. T. Bekolay, "Biologically inspired methods in speech recognition and synthesis: closing the loop by," Ph.D. Thesis, Department of Computer Science, University of Waterloo, 2016.
5. C. Eliasmith, *How to Build a Brain: A Neural Architecture for Biological Cognition*, Oxford University Press, NY, 2013.
6. S. D. Levy and R. Gayler, "Vector symbolic architectures: A new building material for artificial general intelligence," in *Proceedings of the 1st Conference on Artificial General Intelligence*, 2008, pp. 414-418.
7. P. Kanerva, "Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors," *Cognitive Computation*, Vol. 1, 2009, pp. 139-159.

8. G. Montone, J. K. O'Regan, and A. V. Terekhov, "Hyper-dimensional computing for a visual question-answering system that is trainable end-to-end," in *Proceedings of the 31st Conference on Neural Information Processing Systems*, 2017, arXiv:1711.10185.
9. M. S. Gazzaniga, R. B. Ivry, and G. R. Mangun, *Cognitive Neuroscience: The Biology of the Mind*, 4th ed., W. W. Norton & Company, NY, 2014.
10. J. H. McDermott, "Audition," in K. N. Ochsner, and S. M. Kosslyn (eds.), *The Oxford Handbook of Cognitive Neuroscience*, *Vol. 1: Core Topics*, Oxford University Press, NY, 2014, pp. 135-170.
11. D. Oertel and A. J. Doupe, "The auditory central nervous system," in E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, and A. J. Hudspeth (eds.), *Principles of Neural Science*, 5th ed., McGraw-Hill Companies, NY, 2013, pp. 682-711.
12. E. Bigand, C. Delbe, B. Poulin-Charronnat, M. Leman, and B. Tillmann, "Empirical evidence for musical syntax processing? Computer simulations reveal the contribution of auditory short-term memory," *Frontiers in Systems Neuroscience*, Vol. 8, 2014, p. 94.
13. V. Tyagi and C. Wellekens, "On desensitizing the Mel-Cepstrum to spurious spectral components for Robust Speech Recognition," in *Proceedings of IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing*, Vol. 1, 2005, pp. 529-532.
14. C.-Y. Chen, "Navigation of an indoor robot according to estimation of sound arrival direction using a humanoid binaural localization system," Master Thesis, Graduate Institute of Communication Engineering, National Taiwan University, 2018
15. D. P. Kingma and J. Ba, "Adam: A method for Stochastic optimization," *arXiv pre-print arXiv:*1412.6980, 2014.
16. J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech & Language*, Vol. 27, 2013, pp. 621-633.
17. M. M. Breunig, H. P. Kriegel, R. T. NG, and J. Sander, "LOF: Identifying density-based local outliers," in *Proceedings of ACM SIGMOD International Conference on Management of Data*, 2000, pp. 93-104.
18. B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," *Advances in Neural Information Processing Systems*, Vol. 12, 2000, pp. 582-588.

Che-Jui Chang (張哲睿) was born in New Taipei City, Taiwan in 1994. He received B.S. degree in Physics in 2016 and M.S. degree in Communication Engineering in 2018 from National Taiwan University. From 2016 to 2018, he was a Research Assistant in the Cognitive Neuro Robotics Laboratory, National Taiwan University. His research interest includes anomaly detection, acoustic signal processing, machine learning, image processing, and cognitive neurorobotics.

Shyh-Kang Jeng (鄭士康) received B.S. and Ph.D. degrees in Department of Electrical Engineering from National Taiwan University, Taipei, Taiwan, in 1979 and 1983, respectively. In 1981 he joined the faculty of the Department of Electrical Engineering, National Taiwan University, where he is now a Professor. From 1985 to 1993 he visited University of Illinois, Urbana-Champaign, USA, as a Visiting Research Associate Professor and a Visiting Research Professor several times. In 1999 he visited Center for Computer Research in Music and Acoustics, Stanford University, USA, for half of a year. He is also a recipient of 1998 Outstanding Research Award of National Science Council, Taiwan, 2004 Outstanding Teaching Award of National Taiwan University, and 2018 Outstanding Chair Professor of Taiwan Electromagnetic Industry-Academia Consortium. His research interest includes theory and applications of electromagnetics, propagation and signal processing of acoustic waves, computational cognitive neuroscience, machine learning and cognitive neurorobotics.